E3-VITS: Emotional End-to-End TTS with Cross-speaker Style Transfer

Wonbin Jung [1 †] **Junhyeok Lee** [2 †]

Abstract

Since previous emotional TTS models are based on a two-stage pipeline or additional labels, their training process is complex and requires a high labeling cost. To deal with this problem, this paper presents E3-VITS, an end-to-end emotional TTS model that addresses the limitations of existing models. E3-VITS synthesizes high-quality speeches for multi-speaker conditions, supports both reference speech and textual description-based emotional speech synthesis, and enables cross-speaker emotion transfer with a disjoint dataset. To implement E3-VITS, we propose batch-permuted style perturbation, which generates audio samples with unpaired emotion to increase the quality of cross-speaker emotion transfer. Results show that E3-VITS outperforms the baseline model in terms of naturalness, speaker and emotion similarity, and inference speed.

1. Introduction

Over the past few years, the development of neural text-to-speech (TTS) systems has shown significant advancements. However, the majority of studies have primarily focused on synthesizing speech in a standard reading style. This has led to the emergence of enhancing the expressiveness of generated speech as a new challenge. A potential method to address this challenge is to consider various conditions, such as prosody (Ren et al., 2019; 2021; Łańcucki, 2021; Bak et al., 2021), language (Cho et al., 2022; Casanova et al., 2022), and emotion (Wang et al., 2018). To achieve a more human-like TTS system, it is crucial to improve the modeling of varying emotions (Tan et al., 2021).

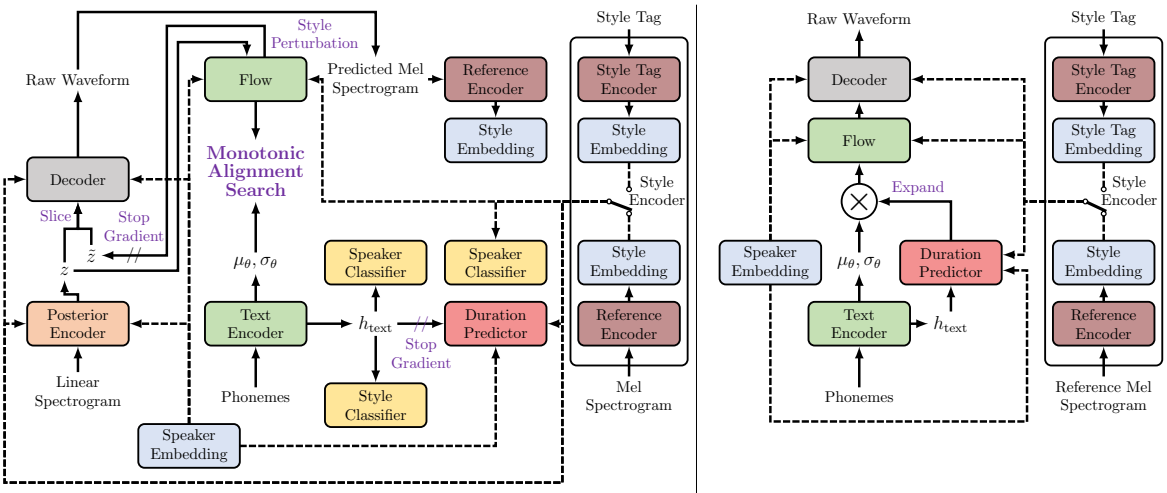
The most naive approach to implementing emotional TTS models is relying on supervised learning (Lee et al., 2017; Liu et al., 2021; Lei et al., 2021) with labeled emotional speech corpora. However, this requires a large speech corpus for each pair of every speaker and emotion. Cross-speaker emotion transferable models (Xue et al., 2021; Kulkarni et al., 2021; An et al., 2021; Sam Ribeiro et al., 2022; Terashima et al., 2022) alleviate this limitation but still limited to predefined categorical labels. To address this problem, unsupervised learning-based models (Wang et al., 2018; Qiang et al., 2022) utilize a reference encoder (Skerry-Ryan et al., 2018). In these models, the reference encoder extracts emotional features from a reference speech, allowing for the free use of diverse emotions.

To combine the advantages of both categorical labels and reference encoder, some previous studies (Kim et al., 2021b; Shin et al., 2022) have introduced certain language models (Reimers & Gurevych, 2019; Brown et al., 2020). These models employ both reference encoder and language model-based style tag encoder, which extracts a style embedding from a textual description of a speaking style called a style tag. By leveraging linguistic information, these models are able to synthesize speech with both seen and unseen style tags. However, these models are two-stage models that require sequential training or fine-tuning of the vocoder, resulting in an increased effort compared to single-stage training. While there have been several end-to-end TTS models with single-stage training, their application in emotional speech synthesis has been relatively limited. For instance, Period VITS (Shirahata et al., 2022) is an end-to-end emotional TTS model that utilizes pitch to achieve stability. However, this model relies on external modules and manually labeled phoneme duration from professional annotators, and does not support cross-speaker emotion transfer.

In this study, we present E3-VITS, a multi-speaker end-to-end emotional TTS model that facilitates speech synthesis using both reference speech and textual emotional descriptions, without the constraint of predefined categorical labels. We adopt domain adversarial training (DAT) to disentangle text representation from speaker and emotion features. To enhance the expressiveness of cross-speaker emotion transfer, we introduce batch-permuted style perturbation. Using these techniques, our model generates high-quality emotional speeches without any external modules or additional duration data. By allowing cross-speaker emotion transfer with a disjoint dataset, E3-VITS reduces the demand for large quantities of emotional speech corpora, enabling more cost-effective training.

† Work done at maum.ai Inc. [1]Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea [2]Supertone Inc.. Correspondence to: Wonbin Jung <santabin@kaist.ac.kr>.

Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023.



(a) Training procedure (b) Inference procedure

Figure 1. Block diagram of a system overview of E3-VITS in (a) training procedure and (b) inference procedure.

2. Method

The proposed E3-VITS model is built upon the architecture of VITS (Kim et al., 2021a) as illustrated in Figure 1. We add style embedding and the style encoder (Shin et al., 2022) to enable emotion control and attach domain classifiers to prevent entanglement. We also modify generative adversarial network training with batch-permuted style perturbation to improve the quality of cross-speaker emotion transfer. To ensure stable duration prediction (Casanova et al., 2022), the stochastic duration predictor of VITS is replaced with the deterministic duration predictor of FastSpeech 2 (Ren et al., 2021). This replacement improves the robustness of the model and ensures accurate phoneme duration prediction.

2.1. Style Embedding

In E3-VITS, emotional information is incorporated by utilizing a style embedding, which has the same dimension of 256 as the speaker embedding. The style embedding is obtained from the style encoder, which is composed of the reference encoder (Min et al., 2021) and SBERT (Reimers & Gurevych, 2019)-based style tag encoder (Shin et al., 2022). The style embedding is inputted into each module alongside the speaker embedding. In each training iteration, E3-VITS randomly selects a style embedding from the reference speech or the style tag with an equal probability.

2.2. Domain Adversarial Training

To separate speaker information from text representation h_{text}, we implement DAT (Ganin et al., 2016) which is used in TTS (Cho et al., 2022; Li et al., 2022; Kim & Chang, 2022) by integrating a speaker reversal classifier. We further introduce a style reversal classifier that classifies the style tag to remove emotional information from the text representation. However, in our preliminary experiments, the style classifier does not exhibit significant improvement due to overfitting. To avoid the complexity inherent in style tags, which are fine-grained labels, we employ more general classes, emotional categories, as coarse-grained labels in the style classification, which lead to more expressive synthesized speech. These two labels are detailed in Section 3.1. Both the speaker and style classifiers consist of fully connected layers and a gradient reversal layer, receiving the hidden text representation h_{text} as an input. Additionally, we include the same speaker classifier that gets the style embedding from the style encoder as input to prevent the potential leakage of speaker information.

2.3. Batch-permuted Style Perturbation

Inspired by the adversarial loss of X. An *et al*. (An et al., 2021), we introduce a method to improve the quality of the synthesized speech with unpaired speaker and emotion that are not paired in the dataset by feeding not only generated audio sample with paired speaker and emotion, but also generated audio sample with unpaired speaker and emotion set during training. In the proposed method, we perturb the style embedding of latent variables z in a batch by referring voice conversion process of VITS through its flow module (Kim et al., 2020; 2021a). Thus, style-perturbed latent variables $\tilde{z}$ of the unpaired set are calculated as follows:

$$\tilde{z} = f\left(f^{-1}(z|s,e)|s,\tilde{e}\right), \tag{1}$$

where f denotes normalizing flow, z are latent variables of the paired set, s and e are paired speaker embedding and style embedding, and $\tilde{e}$ is a style embedding that is not paired with the speaker of s in the training dataset. For each batch, unpaired style embeddings are generated from randomly permuted paired style embeddings. A stop gradient operator (van den Oord et al., 2017) is applied after the forward process of flow to limit its impact on other modules. After calculating $\tilde{z}$, z and $\tilde{z}$ are concatenated and fed into the decoder. In the discriminator, the paired ground truth audio

sample is utilized as a real sample for the generated audio sample of unpaired set. Hence, discriminator adversarial loss $\mathcal{L}_{adv}^{D}$ and generator adversarial loss $\mathcal{L}_{adv}^{G}$ of E3-VITS are described as:

$$\begin{aligned}\mathcal{L}_{adv}^{D} =&\mathbb{E}_{y,z}\left[(D(y)-1)^2+D\left(G(z)\right)^2\right]\\&+\mathbb{E}_{y,\tilde{z}}\left[(D(y)-1)^2+D\left(G(\tilde{z})\right)^2\right],\end{aligned}\tag{2}$$

$$\begin{aligned}\mathcal{L}_{adv}^{G} =&\mathbb{E}_{z}\left[(D(G(z))-1)^2\right]\\&+\mathbb{E}_{\tilde{z}}\left[(D(G(\tilde{z}))-1)^2\right],\end{aligned}\tag{3}$$

where G denotes the decoder, D denotes the discriminator, and y is the ground truth waveform. Since feature matching loss (Larsen et al., 2016) is used in adversarial training of VITS, we also modify feature matching loss $\mathcal{L}_{fm}^{G}$ to consider the style perturbation:

$$\begin{aligned}\mathcal{L}_{fm}^{G} =&\mathbb{E}_{y,z}\left[\sum_{l=1}^{T}\frac{1}{N_l}\left\|D^l(y)-D^l(G(z))\right\|\right]\\&+\mathbb{E}_{y,\tilde{z}}\left[\sum_{l=1}^{T}\frac{1}{N_l}\left\|D^l(y)-D^l(G(\tilde{z}))\right\|\right],\end{aligned}\tag{4}$$

where T is the total number of layers of the discriminator, and $D^l(x)$ denotes outputted feature map of N_l features from the lth layer of the discriminator with input x.

2.4. Objective Function

To learn emotional features, we employ the loss function $\mathcal{L}_{style}$ comprises of the three style losses (Shin et al., 2022) given by:

$$\mathcal{L}_{style}=\alpha\mathcal{L}_{s}^{emb}+\beta\mathcal{L}_{s}^{rec}+\gamma\mathcal{L}_{s}^{con},\tag{5}$$

where $\mathcal{L}_{s}^{emb}$, $\mathcal{L}_{s}^{rec}$, and $\mathcal{L}_{s}^{con}$ denotes style embedding loss, style reconstruction loss, and style contrastive loss, respectively with loss weights $\alpha=\beta=45$ and $\gamma=0.45$. The ratio of the weights is set consistently with the setting of Y. Shin *et al*. (Shin et al., 2022), while weights are scaled up to the same weight as mel reconstruction loss to integrate the style encoder with VITS. The temperature of L_{s}^{con} is fixed at 1.0 to stabilize training.

Thus, the total training objective functions of discriminator $\mathcal{L}_{total}^{D}$ and generator $\mathcal{L}_{total}^{G}$ are defined as:

$$\mathcal{L}_{total}^{D} =\mathcal{L}_{adv}^{D},\tag{6}$$

$$\begin{aligned}\mathcal{L}_{total}^{G} =&\mathcal{L}_{recon}+\mathcal{L}_{kl}+\mathcal{L}_{dur}\\&+\mathcal{L}_{adv}^{G}+\mathcal{L}_{fm}^{G}+\mathcal{L}_{style}+\lambda\mathcal{L}_{dat},\end{aligned}\tag{7}$$

where $\mathcal{L}_{recon}$, $\mathcal{L}_{kl}$, $\mathcal{L}_{dur}$ comes from losses for training original VITS, $\mathcal{L}_{dat}$ represents total loss for classification loss in DAT, and λ is the scale factor. $\mathcal{L}_{dat}$ is the sum of the three reversal classification losses with equal weights and λ follows the schedule in SANE-TTS (Cho et al., 2022).

3. Experiments

3.1. Data

In this study, we utilize FSNR0 (Kim et al., 2021b) Korean Style Tagging TTS dataset[1] which includes speech recordings, transcriptions, emotional category, and style tags from 8 professional voice actors. The style tags are concise descriptions of the speaking style with a total of 332 tags classified into 16 general emotional categories. Examples of style tags and their corresponding emotional categories are shown in Table 1. To create a neutral source dataset, we select 2 speakers from the dataset and only use samples with a neutral style tag for these speakers. Data preprocessing involves the conversion of transcriptions into phoneme sequences using an internal grapheme-to-phoneme converter and the resampling to 22.05 kHz. The dataset is partitioned into training and evaluation sets comprising 115,227 and 480 samples, respectively.

Table 1. Examples of style tags and their emotional categories. We also add their English translation for understanding.

Emotional category	Style tag	Translation
ANGRY	위협하듯	threateningly
ANGRY	심술난듯	nasty
SAD	미안한듯	sorry
JOY	농담하듯	jokingly
JOY	신난듯	excitedly

3.2. Experimental Setup

For the language model of the style tag encoder, we utilize an implementation of Korean SBERT[2] (Reimers & Gurevych, 2019). Since the VITS architecture feeds sliced latent variables into the decoder, we set the length of the sliced segment to 16,384 samples, twice that of VITS. This is done to provide the reference encoder with sufficient linguistic information when it receives the predicted mel spectrogram from the decoder. We train our proposed model for 150 epochs using two NVIDIA A100 GPUs with an initial learning rate of 2.0×10^{-5} and a batch size of 48, employing mixed precision and gradient clipping during the training process. The other hyperparameters follow the original VITS and style encoder. As a baseline model, we compare E3-VITS to Tacotron 2-GST[3] (Wang et al., 2018) with HiFi-GAN[4] (Kong et al., 2020) vocoder which is also used as a baseline in previous studies (Kim et al., 2021b; Shin et al., 2022) of language model-based emotional TTS.

[1]https://www.aihub.or.kr/

[2]https://github.com/BM-K/Sentence-Embedding-Is-All-You-Need

[3]https://github.com/jinhan/tacotron2-gst

[4]https://github.com/jik876/hifi-gan

Table 2. Results of naturalness, speaker similarity, and emotion similarity MOS evaluation

Method	Naturalness			Speaker similarity			Emotion similarity	
	Neutral	Seen emotional	Unseen emotional	Neutral	Seen emotional	Unseen emotional	Seen emotional	Unseen emotional
Ground truth	4.56 ± 0.11	4.79 ± 0.06	-	4.58 ± 0.13	4.44 ± 0.12	-	4.01 ± 0.14	-
Tacotron 2-GST	1.42 ± 0.13	1.19 ± 0.07	1.20 ± 0.08	1.53 ± 0.16	1.26 ± 0.11	1.47 ± 0.16	1.31 ± 0.12	1.18 ± 0.09
E3-VITS (style tag)	$\mathbf{3.00 \pm 0.16}$	$\mathbf{2.96 \pm 0.14}$	$\mathbf{3.17 \pm 0.17}$	$\mathbf{3.77 \pm 0.19}$	$\mathbf{3.28 \pm 0.22}$	$\mathbf{3.13 \pm 0.21}$	$\mathbf{2.83 \pm 0.22}$	$\mathbf{3.02 \pm 0.23}$
E3-VITS (reference)	$\mathbf{2.92 \pm 0.14}$	$\mathbf{3.04 \pm 0.17}$	$\mathbf{3.03 \pm 0.16}$	$\mathbf{3.62 \pm 0.19}$	$\mathbf{3.17 \pm 0.23}$	$\mathbf{3.27 \pm 0.22}$	$\mathbf{2.85 \pm 0.21}$	$\mathbf{2.98 \pm 0.20}$

3.3. Evaluation Metric

We assess the speech naturalness, speaker similarity, and emotion similarity of three methods - Tacotron 2-GST (baseline), E3-VITS (reference), and E3-VITS (style tag) - using a 5-point mean opinion score (MOS) test on a 95% confidence interval. E3-VITS (reference) and E3-VITS (style tag) are speech synthesis methods that use reference speech and style tag, respectively. We measure the MOS scores separately for neutral and emotional speech synthesis. To validate the cross-speaker emotion transferability, the emotional synthesis is categorized by its speaker. Emotional target and neutral source speakers speak with seen and unseen emotion, respectively.

To generate audio samples for evaluation, we select 2 emotional target speakers whose emotional speeches are in the training dataset. As neutral source speakers, we use 2 speakers in the neutral source dataset, defined in Section 3.1. Then we randomly choose 10 emotional and 5 neutral samples of each speaker from the evaluation set. But for neutral source speakers, we gather emotional samples from the portion that we subtracted from the FSNR0 dataset to create the neutral source dataset. We synthesize speeches with the corresponding text and speaker of the chosen ground truth samples via the three speech synthesis methods. Tacotron 2-GST (baseline) and E3-VITS (reference) provide emotion from the chosen samples as reference speeches, while E3-VITS (style tag) uses style tags of the samples. Raters grade ground truth and synthesized samples to assess speech naturalness. To evaluate speaker and emotion similarity, raters compare the synthesized samples to the ground truth samples with the same speaker and style tag. Every sample and sample pair is rated by 6 native Korean speakers.

The inference speed of E3-VITS is compared with the baseline by evaluating the real-time factors (RTF). RTF represents the time taken in seconds by the system to generate a raw waveform of a second. To calculate RTF, we randomly select 50 utterances from the evaluation set and divide the total processing time by the length of synthesized audio samples. The experiments are conducted on a single NVIDIA A100 GPU with a batch size of 1.

4. Results and Discussion

Audio samples of E3-VITS are available on demo page[5].

4.1. Speech Quality

Table 2 presents the naturalness MOS scores for both ground truth and synthesized samples using various methods. The results show that E3-VITS-based approaches outperform Tacotron 2-GST in both neutral and emotional speech synthesis. We found that Tacotron 2-GST tends to produce samples with unclear pronunciation and unnatural duration in some cases, caused by attention error. In contrast, E3-VITS avoids these issues by utilizing the monotonic alignment search (Kim et al., 2020) for duration prediction, generating clearer speeches. Our proposed model does not exhibit a significant difference in performance between neutral and emotional speech synthesis nor between unseen and seen speech synthesis, suggesting that the model performs equally well across different scenarios.

The results in Table 2 indicate that E3-VITS is more effective than the baseline in preserving speaker characteristics. This is likely due to the distortion of speaker characteristics resulting from degraded speech quality in the baseline model. Notably, E3-VITS exhibits higher speaker similarity MOS in neutral synthesis compared to emotional synthesis, which is attributed to the difficulty of maintaining speaker identity in the presence of a broad range of emotions.

Table 2 reveals that E3-VITS achieves higher emotion similarity MOS than Tacotron 2-GST in the emotion speech synthesis. Moreover, both methods with style tag and reference speech show similar degrees of scores, regardless of whether a seen or unseen emotion is used.

E3-VITS (reference) and E3-VITS (style tag) do not show significant differences in every metric, meaning that style tags function as conditions as well as reference speeches.

[5]https://wonbin-jung.github.io/e3-vits/

4.2. Inference Speed

Table 3. Comparison of inference speed

Method	Inference speed (RTF)	Inference speedup
Tacotron 2-GST	2.01×10^{-1}	-
E3-VITS (style tag)	$\mathbf{1.98 \times 10^{-2}}$	10.2 ×
E3-VITS (reference)	$\mathbf{1.77 \times 10^{-2}}$	11.4 ×

As presented in Table 3, E3-VITS outperforms the Tacotron 2-GST in terms of the inference speed. Inference with both style tag and reference speech is more than 10 times faster than inference with the baseline model. Since E3-VITS is single-staged and non-autoregressive, it overcomes the two-stage autoregressive baseline model.

5. Conclusions

This paper presents E3-VITS, an end-to-end emotional TTS model capable of synthesizing speech using reference speech and textual description, without predefined categorical labels. The model utilizes domain adversarial training to separate text representation from speaker and emotion features, and employs style perturbation to improve the expressiveness of cross-speaker emotion transfer. The proposed model outperforms the baseline in terms of naturalness, speaker and emotion similarity, and inference speed. Additionally, it reduces the need for labeled emotional data, enabling cross-speaker emotion transfer. Future research directions include leveraging pitch information for enhanced expressiveness and exploring automatic style tag labeling for datasets without style tags.

References

An, X., Soong, F. K., and Xie, L. Improving Performance of Seen and Unseen Speech Style Transfer in End-to-End Neural TTS. In *INTERSPEECH*, pp. 4688–4692, 2021.

Bak, T., Bae, J.-S., Bae, H., Kim, Y.-I., and Cho, H.-Y. FastPitchFormant: Source-Filter Based Decomposed Modeling for Speech Synthesis. In *INTERSPEECH*, pp. 116–120, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. In *NeurIPS*, volume 33, pp. 1877–1901, 2020.

Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E., and Ponti, M. A. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone. In *ICML*, volume 162, pp. 2709–2720, 2022.

Cho, H., Jung, W., Lee, J., and Woo, S. H. SANE-TTS: Stable And Natural End-to-End Multilingual Text-to-Speech. In *INTERSPEECH*, pp. 1–5, 2022.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Kim, J., Kim, S., Kong, J., and Yoon, S. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. In *NeurIPS*, volume 33, pp. 8067–8077, 2020.

Kim, J., Kong, J., and Son, J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *ICML*, volume 139, pp. 5530–5540, 2021a.

Kim, M., Cheon, S. J., Choi, B. J., Kim, J. J., and Kim, N. S. Expressive Text-to-Speech Using Style Tag. In *INTERSPEECH*, pp. 4663–4667, 2021b.

Kim, M.-K. and Chang, J.-H. Adversarial and Sequential Training for Cross-lingual Prosody Transfer TTS. In *INTERSPEECH*, pp. 4556–4560, 2022.

Kong, J., Kim, J., and Bae, J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *NeurIPS*, volume 33, pp. 17022–17033, 2020.

Kulkarni, A., Colotte, V., and Jouvet, D. Improving transfer of expressivity for end-to-end multispeaker text-to-speech synthesis. In *EUSIPCO*, pp. 31–35, 2021.

Łańcucki, A. Fastpitch: Parallel Text-to-Speech with Pitch Prediction. In *ICASSP*, pp. 6588–6592, 2021.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, volume 48, pp. 1558–1566, 20–22 Jun 2016.

Lee, Y., Rabiee, A., and Lee, S.-Y. Emotional End-to-End Neural Speech Synthesizer. *arXiv preprint arXiv:2206.13404*, 2017.

Lei, Y., Yang, S., and Xie, L. Fine-Grained Emotion Strength Transfer, Control and Prediction for Emotional Speech Synthesis. In *SLT*, pp. 423–430, 2021.

Li, T., Wang, X., Xie, Q., Wang, Z., and Xie, L. Cross-Speaker Emotion Disentangling and Transfer for End-to-End Speech Synthesis. *TASLP*, 30:1448–1460, 2022.

Liu, R., Sisman, B., Gao, G., and Li, H. Expressive TTS Training With Frame and Style Reconstruction Loss. *TASLP*, 29:1806–1818, 2021.

Min, D., Lee, D. B., Yang, E., and Hwang, S. J. Meta-StyleSpeech : Multi-Speaker Adaptive Text-to-Speech Generation. In *ICML*, volume 139, pp. 7748–7759, 18–24 Jul 2021.

Qiang, C., Yang, P., Che, H., Wang, X., and Wang, Z. Style-Label-Free: Cross-Speaker Style Transfer by Quantized VAE and Speaker-wise Normalization in Speech Synthesis. In *ISCSLP*, 2022.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP-IJCNLP*, pp. 3982–3992, 2019.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. FastSpeech: Fast, Robust and Controllable Text to Speech. In *NeurIPS*, volume 32, 2019.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *ICLR*, 2021.

Sam Ribeiro, M., Roth, J., Comini, G., Huybrechts, G., Gabryś, A., and Lorenzo-Trueba, J. Cross-Speaker Style Transfer for Text-to-Speech Using Data Augmentation. In *ICASSP*, pp. 6797–6801, 2022.

Shin, Y., Lee, Y., Jo, S., Hwang, Y., and Kim, T. Text-driven Emotional Style Control and Cross-speaker Style Transfer in Neural TTS. In *INTERSPEECH*, pp. 2313–2317, 2022.

Shirahata, Y., Yamamoto, R., Song, E., Terashima, R., Kim, J.-M., and Tachibana, K. Period VITS: Variational Inference with Explicit Pitch Modeling for End-to-end Emotional Speech Synthesis. *arXiv preprint arXiv:2210.15964*, 2022.

Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R., Clark, R., and Saurous, R. A. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. In *ICML*, volume 80, pp. 4693–4702, 10–15 Jul 2018.

Tan, X., Qin, T., Soong, F. K., and Liu, T. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*, 2021.

Terashima, R., Yamamoto, R., Song, E., Shirahata, Y., Yoon, H.-W., Kim, J.-M., and Tachibana, K. Cross-Speaker Emotion Transfer for Low-Resource Text-to-Speech Using Non-Parallel Voice Conversion with Pitch-Shift Data Augmentation. In *INTERSPEECH*, pp. 3018–3022, 2022.

van den Oord, A., Vinyals, O., and kavukcuoglu, k. Neural Discrete Representation Learning. In *NIPS*, volume 30, 2017.

Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., and Saurous, R. A. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. In *ICML*, volume 80, pp. 5180–5189, 2018.

Xue, L., Pan, S., He, L., Xie, L., and Soong, F. K. Cycle consistent network for end-to-end style transfer TTS training. *Neural Networks*, 140:223–236, 2021. ISSN 0893-6080.